A APPENDIX

A.1 DATASET STATISTICS

We detail the dataset statistics for the three benchmark long-tailed recognition datasets in Table 4.

Table 4: Statistics for training data in CIFAR10-LT, CIFAR100-LT and ImageNet-LT.

Dataset	Attribute	Many	Medium	Few	All
CIFAR10-LT (Imba 200)	Classes	7	3	0	10
	Samples	11052	151	0	11203
CIFAR10-LT (Imba 100)	Classes	8	2	0	10
	Samples	12273	133	0	12406
CIFAR10-LT (Imba 10)	Classes	10	0	0	10
	Samples	20431	0	0	20431
CIFAR100-LT (Imba 200)	Classes	31	30	39	100
	Samples	7753	1445	304	9502
CIFAR100-LT (Imba 100)	Classes	35	35	30	100
	Samples	8824	1718	305	10847
CIFAR100-LT (Imba 10)	Classes	70	30	0	100
	Samples	17743	2130	0	19573
ImageNet-LT	Classes	391	473	136	1,000
	Samples	89,293	24,910	1,643	115,846

A.2 ESTIMATION OF INTRA-CLASS VARIANCE

Estimating the covariance matrix from sample data is a non-trivial problem. In this work, we choose the empirical estimator of sample covariance as described in Eq1, which is the maximum likelihood estimator. However, alternate estimators such as the Ledoit-Wolf Ledoit & Wolf (2004) and Oracle Approximating Shrinkage Chen et al. (2010) estimators are also commonly used to estimate covariance. We did not find any difference in our results due to the choice of estimator; the intra-class variance in Eq2 is always negatively correlated to the class frequency.

In Figure7, we plot the intra-class variance for long-tailed CIFAR100-LT with imbalance ratios $\beta \in [200, 100, 10]$. Since this is a more fine-grained dataset with semantic overlap between classes of the sort orchids and poppies or bicycle and motorcycle, the representations of various classes can overlap and affect the estimation of intra-class variance. Therefore, we also consider the 20 superclasses of CIFAR100 to construct CIFAR20-LT, which is coarse-grained and lacks semantic overlap. In Figure8 we plot the the intra-class variance for long-tailed CIFAR20-LT for the various imbalance ratios. Combined, Figure7 and Figure8 indicate that the negative correlation of intra-class variance to class frequency is not dataset specific and is a more general phenomenon. The high degree of variation in the intra-class variance estimate is attributed to (i) the inverse scaling of the MSE in variance estimation to the class frequency, and (ii) the semantic overlap between various classes due to which intra-class variance is not purely class-conditional.

A.3 LEARNING DISTANCE METRIC AND NCM JOINTLY

Beyond learning just class centroids in Learned NCM, we investigated learning the Mahalanobis distance metric in Eq 3 jointly with the centroids. More precisely, we learn the matrix W which parameterizes the Mahalanobis distance. We experimented with three strategies: (i) Global distance metric shared by all classes, and (ii) Local distance metrics, corresponding to a class-conditional matrix W_y for each class y. Our results for long-tailed CIFAR10-LT and CIFAR100-LT (imba 200 for both) are summarized in Table 5.

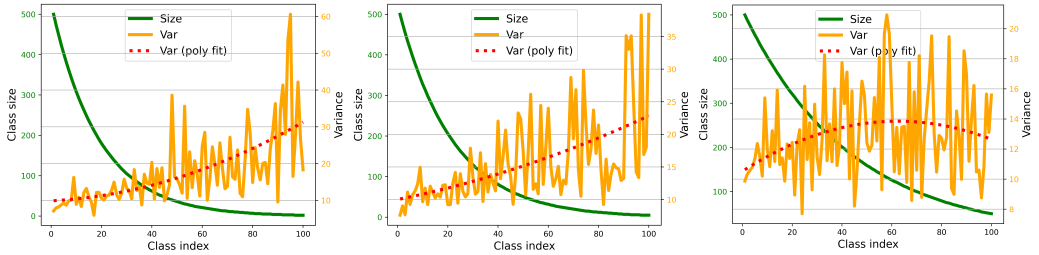

Figure 7: Intra-class variance of representations vs class frequency for long-tailed CIFAR100-LT dataset. Left, middle and right correspond to imbalance ratios of 200, 100 and 10 respectively. Variance is negatively correlated to class frequency.

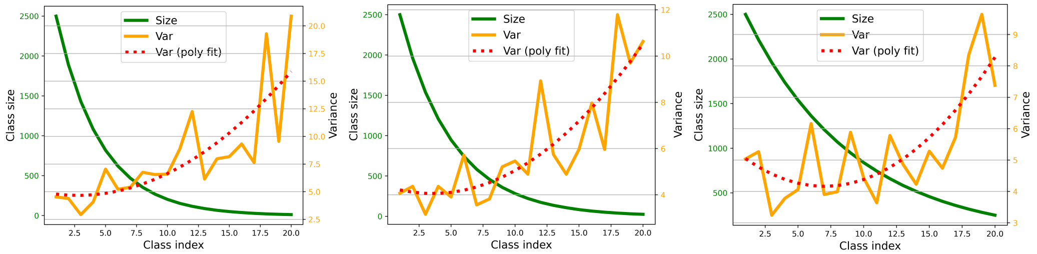

Figure 8: Intra-class variance of representations vs class frequency for long-tailed CIFAR20-LT dataset, consisting of 20 superclasses from the CIFAR100 dataset. Left, middle and right correspond to imbalance ratios of 200, 100 and 10 respectively. Variance is negatively correlated to class frequency.

The results indicate that the learned distance metric only improves *Many* accuracy and in all other cases underperforms Learned NCM. This suggests that Learned NCM is sensitive to choice of distance metric, and keeping the Euclidean distance metric leads to the best results for Learned NCM.

Table 5: Comparision of jointly learned distance metrics and NCM on long-tailed CIFAR10-LT and CIFAR100-LT with imbalance ratio 200.

Dataset	CIFAR10-LT			CIFAR100-LT			
Method	All	Many	Med	All	Many	Med	Few
NCM	79.7	82.2	73.8	43.4	64.6	50.4	21.6
Learned NCM	**80.8**	82.2	**77.6**	**45.7**	66.2	**52.4**	**24.6**
Global Metric + Learned NCM	76.6	**83.1**	61.3	43.4	69.7	48.8	18.9
Local Metric + Learned NCM	80.2	82.1	75.9	43.8	**69.9**	51.0	18.0

A.4 Effect of batch normalization

Batch normalization uses running estimates of the mean and standard deviations statistics to normalize intermediate activations for deep models. For two-stage models used in long-tailed recognition, the batch statistics are used only in stage 1 and after that are kept fixed. However, during training the representations are evolving and so are the batch statistics. Therefore, we experiment with *posthoc* running estimates of mean and standard deviations in the second stage. Since the neural network parameters θ are fixed, the estimates are more precise and can moreover alleviate the biased representations issue discussed in Section 3.

The results are detailed in Table 6. We observe gain in *Few* accuracy due to BN in both Learned NCM and Multi NCM, and gain in *All* accuracy as well. This aligns with our intuition that proper batch normalization can mitigate representation bias in LTR and points to future research directions.

Table 6: Results on the long-tailed CIFAR10-LT and CIFAR100-LT dataset. BN indicates we use posthoc running estimates of mean and standard deviation for the batchnorm layer.

Dataset	CIFAR10-LT			CIFAR100-LT			
Method	All	Many	Medium	All	Many	Medium	Few
Learned NCM	80.8	82.2	77.6	45.7	66.2	52.4	24.6
Learned NCM (BN)	79.7	78.2	**83.2**	45.6	63.4	**52.6**	**26.5**
Multi-NCM	80.8	**82.2**	77.6	45.7	**67.3**	51.8	24.2
Multi-NCM (BN)	**81.4**	81.1	82.2	**45.8**	65.3	52.1	25.8

A.5 DETAILED RESULTS ON CIFAR10-LT

Table 7: Extended results on the long-tailed CIFAR10-LT dataset.

	Imba 200			Imba 100			Imba 10
Method	All	Many	Medium	All	Many	Medium	All
Softmax	74	**83**	52.9	80.3	**81.7**	74.6	90.3
NCM	79.7	82.2	73.8	79.8	79.0	82.8	91.2
Learned NCM	80.8	82.2	77.6	81.6	81.0	83.9	91.4
+LA	**81.3**	81.8	**79.9**	81.4	79.6	**87.4**	**91.7**
Multi-NCM	80.8	82.2	77.6	**81.6**	81.2	83.1	91.3
+LA	81.1	82.4	77.8	81.4	80.1	87.0	91.6